AFRL-IF-RS-TR-2006-223
Final Technical Report
June 2006

TRANSFORMATIVE PATTERN LEARNING

University of Massachusetts

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2006-223 has been reviewed and is approved for publication

APPROVED: /s/

JOHN SPINA
Project Engineer

FOR THE DIRECTOR: /s/

JOSEPH CAMERA, Chief
Information & Intelligence Exploitation Division
Information Directorate

REPORT DOCUMENTATION PAGE	*Form Approved* **OMB No. 0704-0188**

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Service, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* JUN 2006	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* Sep 01 – Nov 05

4. TITLE AND SUBTITLE TRANSFORMATIVE PATTERN LEARNING	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER FA8750-01-2-0566
	5c. PROGRAM ELEMENT NUMBER 62702F
6. AUTHOR(S) David Jensen	**5d. PROJECT NUMBER** EELD
	5e. TASK NUMBER 01
	5f. WORK UNIT NUMBER 18
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Massachusetts 408 Goodell Building Amherst MA 01003-9333	**8. PERFORMING ORGANIZATION REPORT NUMBER** N/A
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFRL/IFED 525 Brooks Rd Rome NY 13441-4505	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSORING/MONITORING AGENCY REPORT NUMBER AFRL-IF-RS-TR-2006-223

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited. PA# 06-486

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Statistical analysis of relational data is a fundamental and novel problem in machine learning and data mining. Such analysis constructs useful statistical models from data about complex relationships among people, places, things, and events. Supported by this research contract, we uncovered fundamental challenges of statistical learning and inference in relational data, we designed and implemented new languages for expressing deterministic and probabilistic dependencies in such data, we developed new algorithms for learning probabilistic models, we implemented an open-source system for knowledge discovery in relational data containing over 40,000 lines of code that has been downloaded more than 1000 times, and we evaluated the utility of those algorithms by undertaking large and realistic applications.

15. SUBJECT TERMS

Statistical learning, inference, relational data, probabilistic models

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** John Spina
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UL	14	**19b. TELEPONE NUMBER (*Include area code*)**

Table of Contents

Summary

Statistical analysis of relational data is a fundamental and novel problem in machine learning and data mining. Such analysis constructs useful statistical models from data about complex relationships among people, places, things, and events. Supported by this research contract, we uncovered fundamental challenges of statistical learning and inference in relational data, we designed and implemented new languages for expressing deterministic and probabilistic dependencies in such data, we developed new algorithms for learning probabilistic models, we implemented an open-source system for knowledge discovery in relational data containing over 40,000 lines of code that has been downloaded more than 1000 times, and we evaluated the utility of those algorithms by undertaking large and realistic applications.

Introduction

This report details the research and development activities of the Knowledge Discovery Laboratory at the University of Massachusetts Amherst Department of Computer Science that were supported under AFRL contract number F30602-01-2-0566 ("Transformative Pattern Learning"). David Jensen, Associate Professor of Computer Science, the Principal Investigator under the contract, directs the Laboratory. The original contract began in FY2002 and continued through December 2005.

The research and development activities supported under this contract produced several basic research results, new algorithms, and prototype systems in the general area of relational knowledge discovery. Relational knowledge discovery focuses on constructing useful statistical models from data about complex relationships among people, places, things, and events. New developments in this area are vital because of the growing interest in mining huge data sets drawn from the web, telecommunications networks, citation graphs, financial transaction histories, social networks, relational databases, and other traces of the activities of complex adaptive systems.

KDL faculty, postdocs, and students were some of the first researchers in this area, which has grown tremendously in the past five years. Our research draws on concepts and techniques from a wide variety of technical areas, including graphical models, classical statistics, causal modeling and inference, heuristic search, query languages, database theory, complex adaptive systems, graph theory, and social network analysis.

Methods, Assumptions, and Procedures

This research used a variety of basic techniques, primarily focused on experimental evaluation of algorithms (e.g., Neville and Jensen 2004), but also including theoretical derivations (e.g., Neville and Jensen 2006), simulation (e.g., Jensen, Rattigan, and Blau 2003; Jensen, Neville, and Gallagher 2004), and applications (e.g., McGovern, Friedland, Hay, Gallagher, Fast, Neville and Jensen 2003).

Several methodological innovations were made during the course of this work, and they are reported in several research papers supported by the contract (e.g., Jensen and Neville 2002a). We discovered that analyzing algorithms for knowledge discovery in relational data often requires subtly different evaluation methods than have been developed for analyzing algorithms for knowledge discovery in propositional data. Additional details are provided in the section on "Results and Discussion".

Our experimental evaluations used a variety of real and simulated data sets, many of which were developed expressly for work under this contract. These data sets included data drawn from the Internet Movie Database (www.imdb.com), the world wide web pages of several university departments of computer science, and the citation network recorded by the computer science bibliography (DBLP; http://www.informatik.uni-trier.de/~ley/db/).

Results and discussion

Our work produced a variety of results, including new statistical models, a new query language, basic and applied results about a new variety of statistical inference, basic research results on the unique characteristics of statistical inference in relational data, prototype software for relational knowledge discovery that has been downloaded more than 1000 times, and example applications of these techniques and software. Each is detailed below.

Statistical models

Our research and development activities produced several new types of statistical models that are useful for relational knowledge discovery. These included:

- *Relational probability trees* — Classification trees are widely used in the machine learning and data mining communities for modeling propositional data. Recent work has extended this basic paradigm to probability estimation trees. Traditional tree learning algorithms assume that instances in the training data are homogenous and independently distributed. Relational probability trees (RPTs) extend standard probability estimation trees to a relational setting in which data instances are heterogeneous and interdependent. Our algorithm for learning the structure and parameters of an RPT searches over a space of relational features that use aggregation functions (e.g. AVERAGE, MODE, COUNT) to dynamically propositionalize relational data and create binary splits within the RPT. Some of our related work identified a number of statistical biases due to characteristics of relational data such as autocorrelation and degree disparity (see below). The RPT algorithm uses a novel form of randomization test to adjust for these biases. On a variety of relational learning tasks, RPTs built using randomization tests are significantly smaller than other models and achieve equivalent, or better, performance (Neville, Jensen, Friedland, and Hay 2003).

- *Relational Bayesian classifiers* — Relational Bayesian classifiers (RBCs) are a modification of the Simple Bayesian Classifier (SBC) for relational data. There exist several Bayesian classifiers that learn predictive models of relational data, but each uses a different estimation technique for modeling heterogeneous sets of attribute values. We considered four estimation techniques and evaluated them on real-world data sets. The best estimator assumes each multiset value is independently drawn from the same distribution, and it performs remarkably well across a range of data sets (Neville, Jensen, and Gallagher 2003).

- *Relational dependency networks* — Relational dependency networks (RDNs) are a new form of graphical model capable of reasoning about joint probability distributions in relational data. RDNs are particularly strong in their ability to learn and reason with cyclic relational dependencies. We have shown that RDNs can be learned on a number of real-world datasets, and we have evaluated the models in a classification context, showing significant performance improvements over existing joint and conditional probability models (Neville and Jensen 2005).

QGraph

In addition to statistical models, we produced significant extensions and prototype versions of QGraph, a visual query language designed to support relational knowledge discovery. See the technical report (Blau, Immerman, and Jensen, 2002) for a full description of the language.

QGraph has some important differences from standard relational database query languages such as SQL. Most obviously, QGraph is a visual query language. Users create queries by drawing the desired graph structure and adding restrictions in the form of textual conditions, constraints, and numeric annotations. QGraph includes a grouping and counting mechanism that lets users move beyond specifying an exact database structure in a query. Queries can contain generalized structural descriptions, enabling a single query to match a variety of database structures.

Unlike SQL, which returns individual records, QGraph queries return subgraphs, complex objects that correspond to connected sets of objects and relations in the queried database. These subgraphs retain the full details of the database objects and links rather than just providing aggregations such as count or average. Subgraphs give users access to both the content and structure of the matches when analyzing and exploring relational data.

During the period of the contract, we significantly extended the language, and developed key concepts necessary to efficiently implement the language in two different database systems (the standard relational model and a variety of the decomposition storage model). In addition, we produced several prototypes that can process a large subset of the full language.

Collective inference

Procedures for collective inference make simultaneous statistical judgments about the same variables for a set of related data instances. For example, collective inference could be used to simultaneously classify a set of hyperlinked documents or infer the legitimacy of a set of related financial transactions. Recent studies, including several conducted under this contract, show that collective inference can significantly reduce classification error when compared with traditional inference techniques.

We investigated the underlying mechanisms for this error reduction by reviewing past work on collective inference and characterizing different types of statistical models used for making inference in relational data. We showed important differences among these models, and we characterized the necessary and sufficient conditions for reduced classification error based on experiments with real and simulated data. This work can be found in a 2004 paper (Jensen, Neville, and Gallagher 2004). In addition, we developed algorithms for learning and making inferences with a new type of statistical model — a relational dependency network (RDN) (Neville and Jensen 2003, 2004, 2006). These algorithms automatically learn models that can exploit collective inference.

Unique statistical biases

Two common characteristics of relational data sets — concentrated linkage and relational autocorrelation — can cause learning algorithms to be strongly biased toward certain dependencies among variables, irrespective of the predictive power of those dependencies. Informally, concentrated linkage occurs when many objects are linked to a common neighbor, and relational autocorrelation occurs when the values of a given variable are highly uniform among objects that share a common neighbor. We identify these characteristics, define quantitative measures of their severity, and explain how they produce this bias. We showed how linkage and autocorrelation affect a representative algorithm for feature selection by applying the algorithm to both real and synthetic data. This work can be found in a 2002 paper (Jensen and Neville 2002c).

The following year, we discovered that another common characteristic of relational data sets — degree disparity — can lead relational learning algorithms to discover misleading correlations. Degree disparity occurs when the frequency of a relation is correlated with the values of the target variable. In such cases, aggregation functions used by many relational learning algorithms will result in misleading correlations and added complexity in models. We examined this problem through a combination of simulations and experiments. We showed how two novel hypothesis-testing procedures can adjust for the effects of using aggregation functions in the presence of degree disparity. This work can be found in a 2003 paper (Jensen, Neville, and Hay 2003).

Proximity software

During the course of the contract, we released five versions of Proximity, our environment for knowledge discovery in relational data (3.0, 3.1, 4.0, 4.1, and 4.2). Proximity is an open-source software environment, which incorporates a high-performance database, a visual query language (QGraph), methods for automatic construction of statistical models (RPTs, RBCs, and RDNs), a browser-style Interface, methods for importing XML-based data, a Python-based scripting language, and over 200 pages of documentation. Proximity is implemented in over 40,000 lines of Java code, and utilizes Monet, an innovative open-source database. Proximity has been downloaded over 1000 times since its initial release.

Applications

KDL members participated in the 2003 KDD Cup competition, a data mining competition held in conjunction with the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003). The annual competition is the most widely recognized competitive evaluation of technologies and practices of knowledge discovery and data mining. Our entry was awarded first place in the portion of the competition.

For the competition, we analyzed publication patterns in theoretical high-energy physics using a relational learning approach. We focused our analyses on four related areas: understanding and identifying patterns of citations, examining publication patterns at the author level, predicting whether a paper will be accepted by specific journals, and identifying research communities from the citation patterns and paper text. Each of these analyses contributes to an overall understanding of theoretical high-energy physics that could not have been achieved without examining each area in detail.

KDL's team competed in the “open task” (one of four tracks in the 2003 KDD Cup competition) that allowed contestants to define their own analysis tasks within the designated competition data set. A panel of judges selected winners. The title of the UMass entry was “Exploiting Relational Structure to Understand Publication Patterns in High Energy Physics” (McGovern, Friedland, Hay, Gallagher, Fast, Neville and Jensen 2003).

Conclusions

In addition to new languages, algorithms, open-source software, and applications, our research produced a set of more basic conclusions about relational knowledge discovery. These included:

Relational learning improves accuracy

Relational learning greatly expands the set of potential predictors for any given variable in a data set. For example, in data about movies, the success of a movie could depend only on intrinsic characteristics of movies (e.g., genre), but relational methods allow models to consider characteristics of the movie’s studio, actors, director, producer, and characteristics of movies connected to those organizations and people. As a result, relational models can be more accurate.

However, this expanded set of potential statistical dependencies could also increase the variance component of error (with more dependencies to select among, the chance of error increases). In practice, we found that this occurred relatively rarely.

Collective inference improves accuracy

Methods that allowed many relational inferences to be made simultaneously can increase accuracy. We observed this effect in practice (e.g., Neville and Jensen 2003, 2004), and we also showed the deeper causes of this effect (Jensen, Neville, and Gallagher 2004).

New query languages are needed to support relational knowledge discovery

Our early experience with using traditional query languages (e.g., SQL) and our subsequent experience after designing and implementing QGraph indicates that traditional query languages are not well suited to supporting knowledge discovery in relational data. Specifically, traditional query languages do not easy allow users to return portions of the graph. They force returned values to be records with a fixed and rigid structure (e.g., a row in a table). In contrast, we often found it beneficial to return subgraphs that have highly variable structure. On returned subgraph may have a single node and another may have dozens or hundreds of nodes.

Using propositional models on relational data causes statistical biases

Two of our most highly cited pieces of research under this contract concern errors in statistical inference that occur when traditional learning algorithms are applied to relational data. We showed that a common form of correlation in relational data (autocorrelation) and a common form of correlation between attributes and graph structure (degree disparity) can cause serious errors in “naïve” learning algorithms (Jensen and Neville 2002a, 2002c; Jensen, Neville, and Hay 2003).

Models of structure remain elusive

Much of the success of the pattern learning technologies that we have developed over the past four years rest on three types of models — relational Bayesian classifiers, relational probability trees, and relational dependency networks. Each of these models estimates probability distributions of attributes. For example, such models might predict the topic of a meeting or the legitimacy of a financial transaction.

While these models consider the structure of relational data sets during learning and inference, they do not make explicit predictions about either the link structure of data (e.g., inferring that a link exists between two objects) or higher-level group structures in the data (e.g., inferring that persons x, y, and z form a cell). We believe that integrated modeling of attributes, links, and groups are likely to provide impressive gains in accuracy and overall analytic utility.

We have explored link prediction, and we have identified some of the reasons why this task is particularly difficult (Rattigan and Jensen 2005). However, completely integrated models of both structure and attributes are still more dream than reality

Pseudo-likelihood models of relational data can be surprisingly effective

Relational dependency networks, our joint model of attributes in relational data, is referred to as a pseudo-likelihood model. Rather than learn the model that maximizes the joint likelihood of the data, the algorithm constructs the model by maximizing individual conditional likelihoods and then composing those models into a model of the joint probability distribution. This approach seems unlikely to work well, on its face, yet our experiments indicated that such models are surprisingly effective at learning joint models of relational data.

Inferring causality in relational data is a major unexplored area

Our own experience with interpreting relational models is that we often want to interpret them *causally*. That is, we wish to attribute causation to the discovered dependencies, rather than mere correlation. Substantial work on causality has been done in propositional learning, and we suspect that this work could be imported and adapted to perform similar inferences in relational data. However, this is almost completely unexplored.

References

Blau, H., N. Immerman and D. Jensen (2002). A visual language for querying and updating graphs. University of Massachusetts Amherst Computer Science Technical Report 2002-037. For information on QGraph language updates and implementation in Proximity, see the Proximity QGraph Guide.

Jensen, D. and J. Neville (2002a). Autocorrelation and linkage cause bias in evaluation of relational learners. *Proceedings of the 12th International Conference on Inductive Logic Programming*.

Jensen, D. and J. Neville (2002c). Linkage and autocorrelation cause feature selection bias in relational learning. *Proceedings of the 19th International Conference on Machine Learning.*

Jensen, D., J. Neville and B. Gallagher (2004). Why collective inference improves relational classification. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Jensen, D., J. Neville and M. Hay (2003). Avoiding bias when aggregating relational data with degree disparity. *Proceedings of the 20th International Conference on Machine Learning.*

Jensen, D., M. Rattigan and H. Blau (2003). Information awareness: a prospective technical assessment. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

McGovern, A., L. Friedland, M. Hay, B. Gallagher, A. Fast, J. Neville and D. Jensen (2003). Exploiting relational structure to understand publication patterns in high-energy physics. Winning entry: KDD Cup 2003 Open Task, 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Neville, J. and D. Jensen (2003). Collective classification with relational dependency networks. *Proceedings of the 2nd Multi-Relational Data Mining Workshop, 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Neville, J. and D. Jensen (2004). Dependency networks for relational data. *Proceedings of The 4th IEEE International Conference on Data Mining.*

Neville, J. and D. Jensen (2005). Leveraging relational autocorrelation with latent group models. *Proceedings of the 5th IEEE International Conference on Data Mining*.

Neville, J. and D. Jensen (2006). Dependency networks for knowledge discovery in relational data. *Journal of Machine Learning Research (forthcoming).*

Neville, J., D. Jensen and B. Gallagher (2003). Simple estimators for relational Bayesian classifiers. *Proceedings of The 3rd IEEE International Conference on Data Mining.* (Also appeared as University of Massachusetts Amherst, Technical Report 03-04).

Neville, J., D. Jensen, L. Friedland and M. Hay (2003). Learning relational probability trees. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Rattigan, M. and D. Jensen (2005). The case for anomalous link detection. *Proceedings of the 4th Multi-Relational Data Mining Workshop, 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Papers supported by this contract

Blau, H., N. Immerman and D. Jensen (2002). A visual language for querying and updating graphs. University of Massachusetts Amherst Computer Science Technical Report 2002-037. For information on QGraph language updates and implementation in Proximity, see the Proximity QGraph Guide.

Blau, H. and A. McGovern (2003). Categorizing unsupervised relational learning algorithms. *Proceedings of the Workshop on Learning Statistical Models from Relational Data, 18th International Joint Conference on Artificial Intelligence*.

Jensen, D. and J. Neville (2002). Autocorrelation and linkage cause bias in evaluation of relational learners. *Proceedings of the 12th International Conference on Inductive Logic Programming.*

Jensen, D. and J. Neville (2002). Data mining in social networks. National Academy of Sciences Symposium on Dynamic Social Network Analysis.

Jensen, D. and J. Neville (2002). Linkage and autocorrelation cause feature selection bias in relational learning. *Proceedings of the 19th International Conference on Machine Learning*.

Jensen, D. and J. Neville (2002). Schemas and models. *Proceedings of the Multi-Relational Data Mining Workshop, 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Jensen, D., J. Neville and B. Gallagher (2004). Why collective inference improves relational classification. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Jensen, D., J. Neville and M. Hay (2003). Avoiding bias when aggregating relational data with degree disparity. *Proceedings of the 20th International Conference on Machine Learning.*

Jensen, D., J. Neville and M. Rattigan (2003). Randomization tests for relational learning. University of Massachusetts Amherst, Technical Report 03-05.

Jensen, D., M. Rattigan and H. Blau (2003). Information awareness: a prospective technical assessment. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

McGovern, A., L. Friedland, M. Hay, B. Gallagher, A. Fast, J. Neville and D. Jensen (2003). Exploiting relational structure to understand publication patterns in high-energy physics. Winning entry: KDD Cup 2003 Open Task, 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

McGovern, A. and D. Jensen (2003). Identifying predictive structures in relational data using multiple instance learning. *Proceedings of the 20th International Conference on Machine Learning.*

Neville, J., M. Adler and D. Jensen (2004). Spectral clustering with links and attributes. University of Massachusetts Amherst, Technical Report 04-42.

Neville, J. and D. Jensen (2002). Supporting relational knowledge discovery: lessons in architecture and algorithm design. *Proceedings of the Data Mining Lessons Learned Workshop, 19th International Conference on Machine Learning*.

Neville, J. and D. Jensen (2003). Collective classification with relational dependency networks. *Proceedings of the 2nd Multi-Relational Data Mining Workshop, 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Neville, J. and D. Jensen (2004). Dependency networks for relational data. *Proceedings of The 4th IEEE International Conference on Data Mining*.

*Neville, J. and D. Jensen (2006). Dependency networks for knowledge discovery in relational data. *Journal of Machine Learning Research (forthcoming).*

Neville, J., D. Jensen and B. Gallagher (2003). Simple estimators for relational Bayesian classifiers. *Proceedings of The 3rd IEEE International Conference on Data Mining*. (Also appeared as University of Massachusetts Amherst, Technical Report 03-04).

Neville, J., D. Jensen, L. Friedland and M. Hay (2003). Learning relational probability trees. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Neville, J., Ö. Simsek and D. Jensen (2004). Autocorrelation and relational learning: challenges and opportunities. *Proceedings of the Workshop on Statistical Relational Learning, 21st International Conference on Machine Learning*.